Data_clean

Khalid

11/17/2020

```
#I read my dataset from PGA tour data from 2010-2018 about all professional golfers. I used "Clean_name

dirtygolf <- read_csv("raw_data/pgaTourData.csv") %>%
  clean_names() %>%
  mutate(earnings = as.numeric(gsub("," , "" , (str_sub(money, 2 ))))) %>%
  select(-money)
```

```
## Parsed with column specification:
## cols(
##   `Player Name` = col_character(),
##   Rounds = col_double(),
##   `Fairway Percentage` = col_double(),
##   Year = col_double(),
##   `Avg Distance` = col_double(),
##   gir = col_double(),
##   `Average Putts` = col_double(),
##   `Average Scrambling` = col_double(),
##   `Average Score` = col_double(),
##   Points = col_number(),
##   Wins = col_double(),
##   `Top 10` = col_double(),
##   `Average SG Putts` = col_double(),
##   `Average SG Total` = col_double(),
##   `SG:OTT` = col_double(),
##   `SG:APR` = col_double(),
##   `SG:ARG` = col_double(),
##   Money = col_character()
## )
```

```
#This Data set takes the aveage of several golf statistics for many notable players over the years 2010


#I grouped the golfers over the period 2010-2018 in order to average their statistics over the time per

golfers <- dirtygolf %>%
  group_by(player_name) %>%
  summarise(rounds = mean(rounds, na.rm = TRUE),
            fairway_percentage = mean(fairway_percentage, na.rm = TRUE),
            avg_distance = mean(avg_distance, na.rm = TRUE),
            gir = mean(gir, na.rm = TRUE),
            average_putts = mean(average_putts, na.rm = TRUE),
```

```
            average_scrambling = mean(average_scrambling, na.rm = TRUE),
            average_score = mean(average_score, na.rm = TRUE),
            points = mean(points, na.rm = TRUE),
            wins = mean(wins, na.rm = TRUE),
            top_10 = mean(top_10, na.rm = TRUE),
            average_sg_total = mean(average_sg_total, na.rm = TRUE),
            average_putts = mean(average_putts, na.rm = TRUE),
            sg_ott = mean(sg_ott, na.rm = TRUE),
            sg_apr = mean(sg_apr, na.rm = TRUE),
            sg_arg = mean(sg_arg, na.rm = TRUE),
            earnings = mean(earnings, na.rm = TRUE))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
golfers
```

```
## # A tibble: 526 x 16
##    player_name rounds fairway_percent~ avg_distance   gir average_putts
##    <chr>        <dbl>            <dbl>        <dbl> <dbl>         <dbl>
##  1 Aaron Badd~   77.8             53.4         294.  61.9          28.3
##  2 Aaron Watk~   46               63.0         277.  62.2          28.8
##  3 Aaron Wise    90               63.3         303.  68.8          29.2
##  4 Abraham An~   74.5             64.5         286.  64.2          29.0
##  5 Adam Hadwin   94.5             64.4         290.  66.4          28.8
##  6 Adam Schenk   86               57.7         306.  67.9          30.0
##  7 Adam Scott    64.9             60.5         303.  68.9          29.6
##  8 Alex Aragon  NaN              NaN           NaN  NaN           NaN
##  9 Alex Cejka    77.8             66.7         282.  66.1          29.1
## 10 Alex Noren    67               61.7         299.  64.9          29.0
## # ... with 516 more rows, and 10 more variables: average_scrambling <dbl>,
## #   average_score <dbl>, points <dbl>, wins <dbl>, top_10 <dbl>,
## #   average_sg_total <dbl>, sg_ott <dbl>, sg_apr <dbl>, sg_arg <dbl>,
## #   earnings <dbl>
```

```
# I took the top 50 percent of golfers based on the "wins" col
top_golfers <- golfers %>%
  arrange(desc(wins)) %>%
  slice_head(prop = 0.5)

top_golfers
```

```
## # A tibble: 263 x 16
##    player_name rounds fairway_percent~ avg_distance   gir average_putts
##    <chr>        <dbl>            <dbl>        <dbl> <dbl>         <dbl>
##  1 Tiger Woods   66               61.9         298.  67.5          28.8
##  2 Justin Tho~   95.5             56.2         306.  67.7          28.6
##  3 Jordan Spi~   86.8             61.4         294.  66.9          28.3
##  4 Francesco ~   67               68.9         290.  68.4          29.4
##  5 Hunter Mah~   81.4             63.9         294.  67.0          29.4
##  6 Jason Day     72.4             55.6         305.  65.6          28.5
##  7 Jimmy Walk~   85.2             52.3         298.  64.8          28.8
##  8 Martin Kay~   63               61.3         292.  64.7          29.5
```

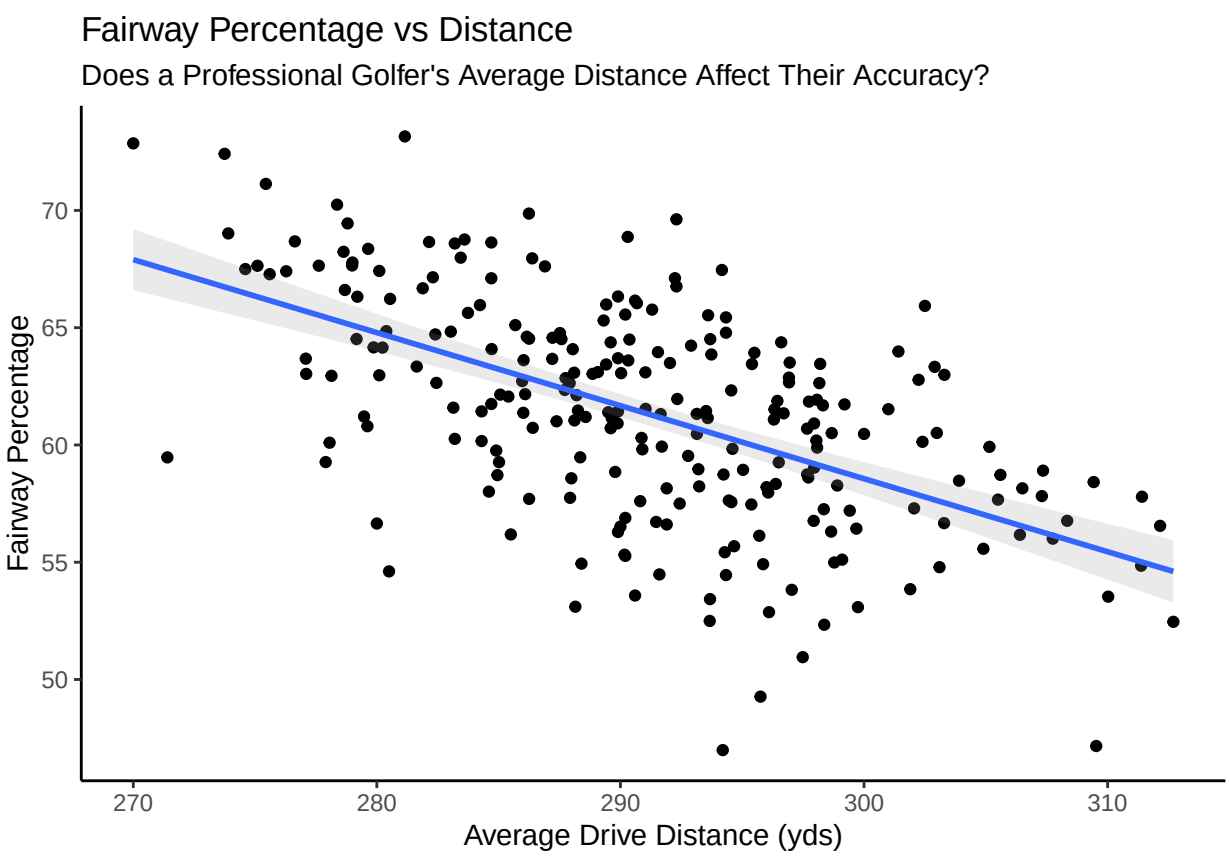
```
##  9 Nick Watney   87              59.9         298.  67.6          29.3
## 10 Patton Kiz~   89.7            54.5         294.  65.7          29.2
## # ... with 253 more rows, and 10 more variables: average_scrambling <dbl>,
## #   average_score <dbl>, points <dbl>, wins <dbl>, top_10 <dbl>,
## #   average_sg_total <dbl>, sg_ott <dbl>, sg_apr <dbl>, sg_arg <dbl>,
## #   earnings <dbl>
```

```
#A scatter plot comparing the average distance off the tee box to the percentage of fairways hit from t
top_golfers%>%
  ggplot(aes(x = avg_distance, y = fairway_percentage))+
  geom_point()+
  theme_classic()+
  geom_smooth(alpha = 0.2, method = "lm")+
  labs(title = "Fairway Percentage vs Distance",
       subtitle = "Does a Professional Golfer's Average Distance Affect Their Accuracy?",
       x = "Average Drive Distance (yds)",
       y = "Fairway Percentage")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 21 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 21 rows containing missing values (geom_point).
```



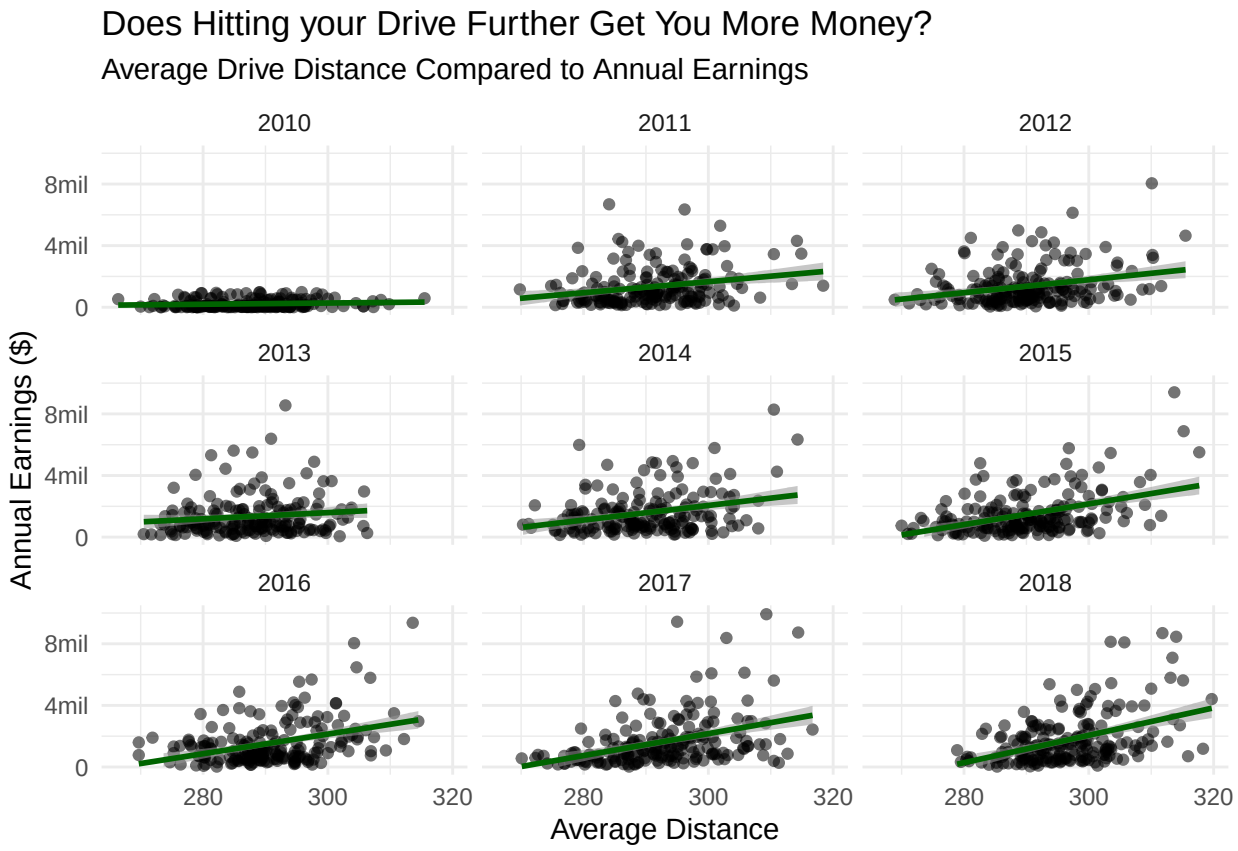
```
#get rid of commas in Funds col *as.numeric*

#a scatterplot comparing several variables to earninigs.
dirtygolf %>%
  ggplot(aes(x = avg_distance, y = earnings, alpha = .2)) +
  geom_point()+
  facet_wrap(~year)+
  theme_minimal()+
  geom_smooth(method = "lm", color = "dark green")+
  theme(legend.position = "none")+
  labs(title = "Does Hitting your Drive Further Get You More Money?",
       subtitle = "Average Drive Distance Compared to Annual Earnings",
       x = "Average Distance",
       y = "Annual Earnings ($)")+
  scale_y_continuous(breaks = c(0,4000000,8000000), labels = c("0", "4mil", "8mil"), limits = c(0,10000
```

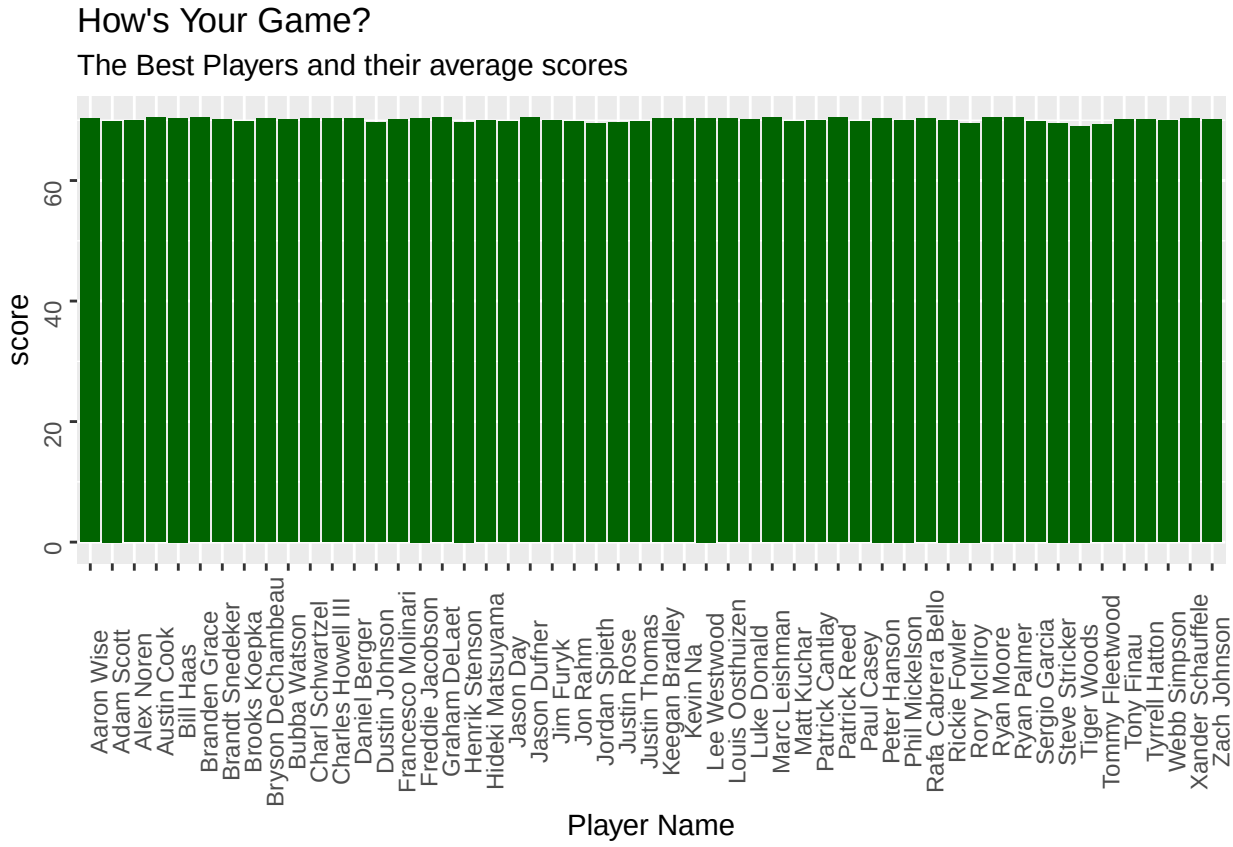
```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 639 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 639 rows containing missing values (geom_point).
```



```
golfers %>%
  arrange(average_score) %>%
  slice_head(prop = 0.1) %>%
  ggplot(aes(x = player_name, y = average_score))+
  geom_col(fill = "dark green")+
  theme(axis.text = element_text(angle = 90))+
  labs(x = "Player Name",
       y = "score",
       title = "How's Your Game?",
       subtitle = "The Best Players and their average scores")
```




```
golf_model <- stan_glm(data = top_golfers,
                       formula = earnings ~ avg_distance,
                       family = gaussian(),
                       refresh = 0)

print(golf_model)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      earnings ~ avg_distance
##  observations: 242
##  predictors:   2
## ------
##               Median      MAD_SD
```

```
## (Intercept)  -15745735.5   2230404.1
## avg_distance     58580.9      7603.7
##
## Auxiliary parameter(s):
##       Median    MAD_SD
## sigma 1046857.9   47292.3
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

~ Rank each player and create comparision data for audience ~get an additional dataset for demographics / sponsors / nationality